Variational Methods for Latent Variable Problems (part 2)

Ryan Giordano (for Johns Hopkins Biostats BLAST working group)

Oct, 2021

Massachusetts Institute of Technology

Outline

Outline for today:

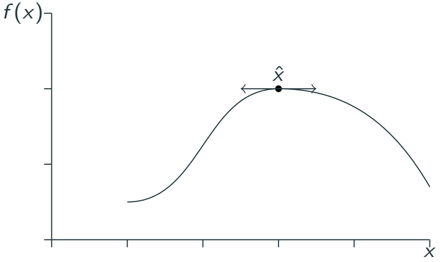
- What counts as variational inference?
- Kullback-Leibler (KL) divergence and "standard" variational inference
- The classical EM algorithm as a special case of variational inference
- Variational inference as a generalization of the EM algorithm
- A quick and incomplete sketch of further topics in variational inference

What counts as variational inference?

Lots of very different procedures go by the name "variational inference." I propose an (idosyncratic) enompassing definition based on the use cases and the name:

Variational inference is inference using optimization.

Think "calculus of variations:" an optimum $\hat{x} = \underset{\theta}{\mathrm{argmax}}\, f(x)$ is characterized by $df/dx|_{\hat{x}} = 0$, i.e. where small variations in $\hat{x}$ result in no changes to the value of $f(\hat{x})$.



By this definition,

- The maximum likelihood estimator (MLE) is VI.
- The Laplace approximation to a Bayesian posterior is VI.
- Markov chain Monte Carlo (MCMC) is not VI.

What counts as variational inference?

A more common definition of VI is the following.

Suppose we have a random variable ξ and a distribution $\mathfrak{p}(\xi)$ that we want to know.

Let y denote data and θ a parameter. Examples:

- The variable is θ, and we wish to know the posterior $\mathfrak{p}(\theta|y)$ (Bayes)
- The variable is y, and we wish to know $\mathfrak{p}(y)$ (MLE)
- The variable is y, and we wish to know the map $\theta \mapsto \mathfrak{p}(y|\theta) = \int p(y, z|\theta)dz$ (marginal MLE)

Let $\mathcal{Q}$ be some class of distributions which may or may not contain $\mathfrak{p}(\xi)$.

Variational inference finds the distribution in $\mathcal{Q}$ closest to $\mathfrak{p}$ according to some measure of "divergence" between distributions:

$$\mathfrak{q}^* = \operatorname*{argmin}_{\mathfrak{q} \in \mathcal{Q}} D(\mathfrak{q}, \mathfrak{p}).$$

The most common choice of "divergence" is the **Kullback-Leibler** (KL) divergence, though other choices are possible (e.g. Li and Turner [2016], Liu and Wang [2016], Ambrogioni et al. [2018]).

KL divergence

The KL divergence is defined as:

$$\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) := \mathop{\mathbb{E}}_{\mathfrak{q}(\xi)}\left[\log \mathfrak{q}(\xi)\right] - \mathop{\mathbb{E}}_{\mathfrak{q}(\xi)}\left[\log \mathfrak{p}(\xi)\right]$$

Some points to be aware of:

- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) \geq 0$
- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) = 0 \Rightarrow \mathfrak{p} = \mathfrak{q}$
- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right) \neq \mathrm{KL}\left(\mathfrak{p}||\mathfrak{q}\right)$
- $\mathrm{KL}\left(\mathfrak{q}||\mathfrak{p}\right)$ is a "strict" measure of closeness [Gibbs and Su, 2002]

Why use KL divergence?

Phony answer: The KL divergence has an information theoretic interpretation [Kullback and Leibler, 1951].

Real answer: Mathematical convenience (normalizing constants pop out).

Example: the MLE minimizes KL divergence. Suppose that $x_n \overset{iid}{\sim} \mathfrak{p}(\cdot)$, and $\mathfrak{q}(\cdot|\theta) \in \mathcal{Q}$ is a (possibly misspecified) parametric family of data distributions. Then

$$\begin{aligned}\hat{\theta} := \underset{\theta}{\mathrm{argmin}}\, \mathrm{KL}\left(\mathfrak{p}||\mathfrak{q}\right) &= \underset{\theta}{\mathrm{argmin}} \left(- \mathop{\mathbb{E}}_{\mathfrak{p}(x_1)}\left[\log \mathfrak{q}(x_1|\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{p}(x_1)}\left[\log \mathfrak{p}(x_1)\right]\right) \\ &= \underset{\theta}{\mathrm{argmax}} \mathop{\mathbb{E}}_{\mathfrak{p}(x_1)}\left[\log \mathfrak{q}(x_1|\theta)\right] \approx \underset{\theta}{\mathrm{argmax}} \frac{1}{N}\sum_{n=1}^{N} \log \mathfrak{q}(x_n|\theta).\end{aligned}$$

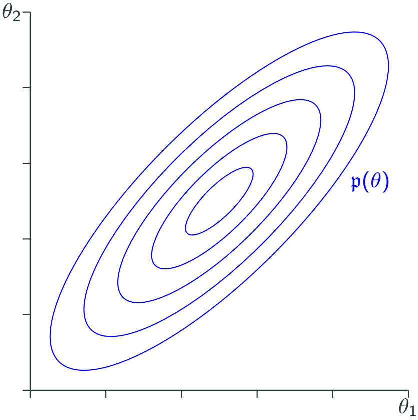
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{All bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



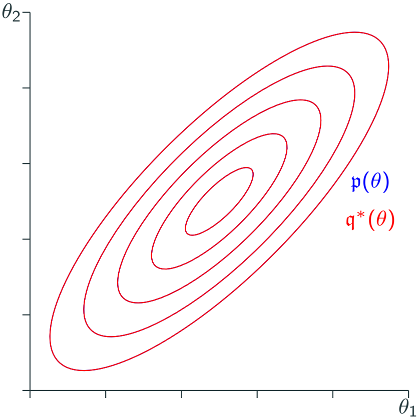
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{All bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Sufficiently expressive families recover the target distribution.

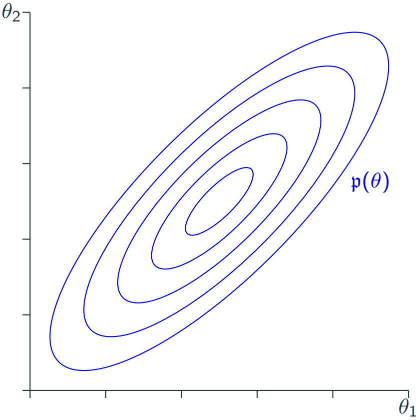
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



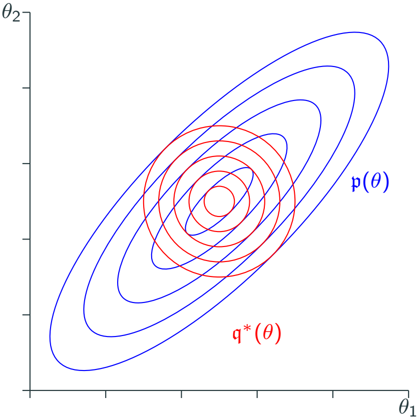
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



KL minimizers "fit inside" the second argument.

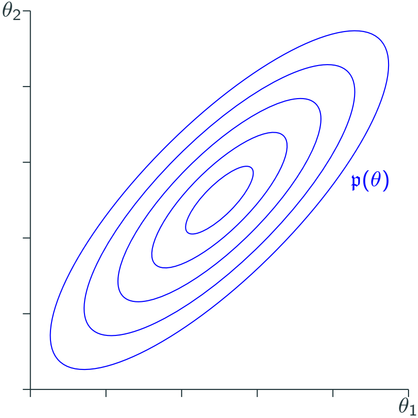
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right) = \\ -\mathop{\mathbb{E}}_{\mathfrak{p}(\theta)}[\log \mathfrak{q}(\theta)] + \mathop{\mathbb{E}}_{\mathfrak{p}(\theta)}[\log \mathfrak{p}(\theta)]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Independent bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \operatorname*{argmin}_{\mathfrak{q}\in\mathcal{Q}} \mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right)$?



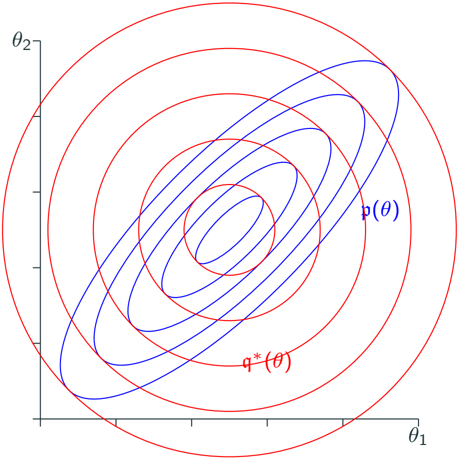
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right) =$$
$$-\underset{\mathfrak{p}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right] + \underset{\mathfrak{p}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right]$$

$\mathfrak{p}(\theta) =$ Correlated bivariate normal

$\mathcal{Q} = \{$Independent bivariate normals$\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\, \mathrm{KL}\left(\mathfrak{p}(\theta)||\mathfrak{q}(\theta)\right)$?



KL minimizers "fit inside" the second argument.

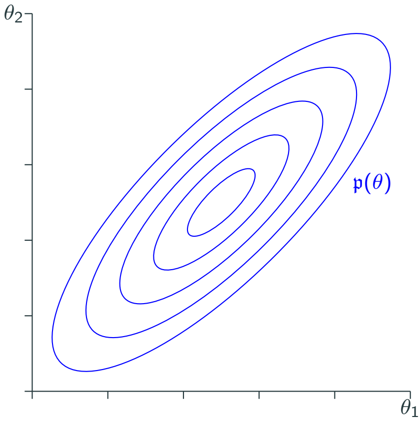
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log\mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log\mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) =$
$$\underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}}\left(-\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log\mathfrak{p}(\theta)\right] + \cancel{\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log\mathfrak{q}(\theta)\right]}\right)?$$



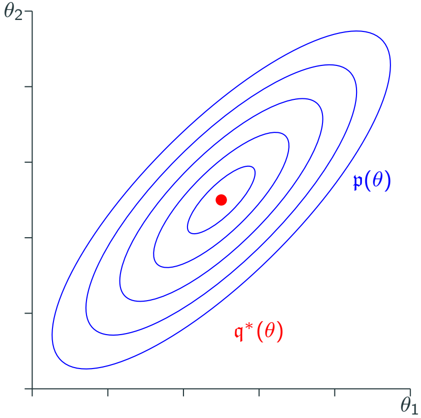
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \left(- \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \cancel{\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]}\right)$?



Without the entropy, the KL minimizer concentrates on the maximum of $\log \mathfrak{p}(\theta)$.

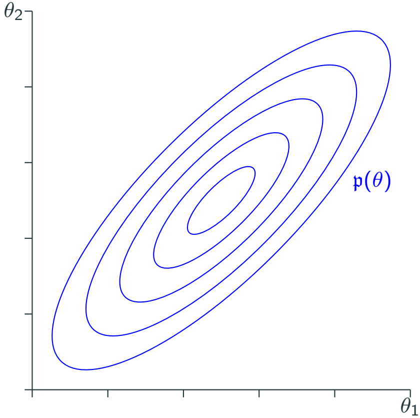
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\left(\cancel{-\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right]} + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]\right)$?



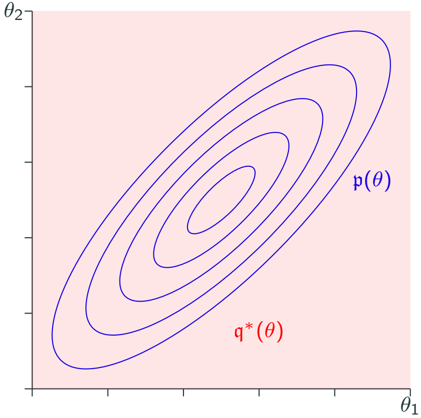
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = -\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{p}(\theta)] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{q}(\theta)]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Bivariate normals}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\mathrm{argmin}}\left(-\cancel{\mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{p}(\theta)]} + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}[\log \mathfrak{q}(\theta)]\right)$?



Without $\log \mathfrak{p}(\theta)$, the KL minimizer is infinitely dispersed.

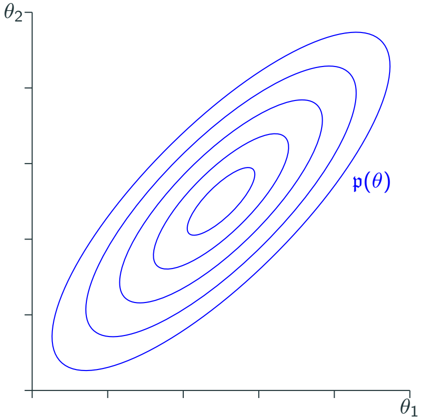
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{Point masses}\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



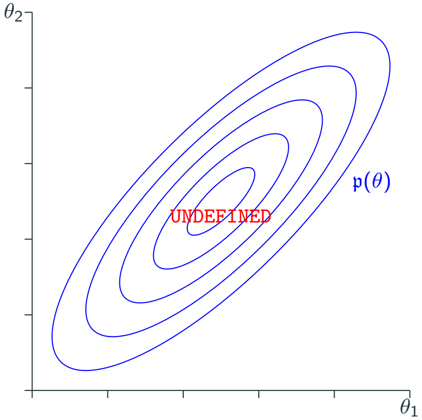
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = - \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{p}(\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(\theta)}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{Point masses}\}$$

What is $\mathfrak{q}^*(\theta) = \operatorname*{argmin}_{\mathfrak{q} \in \mathcal{Q}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Without a common dominating measure, the KL divergence is undefined.

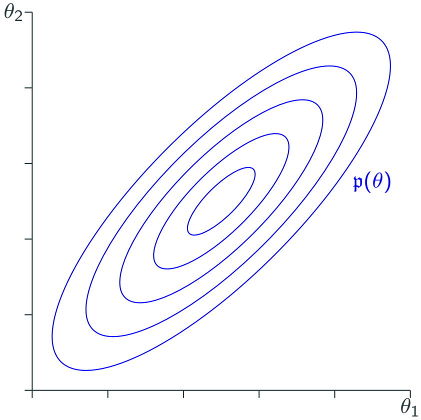
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$

$\mathcal{Q} = \{\text{BVN with small, fixed variance}\}$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



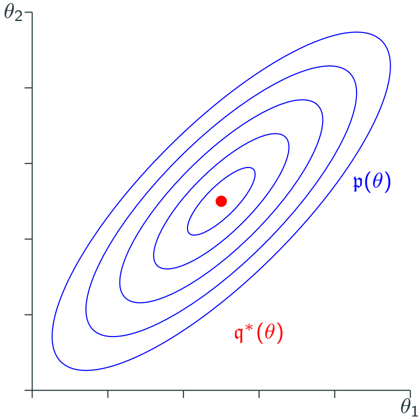
KL divergence exercises

$$\mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right) = \\ -\underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{p}(\theta)\right] + \underset{\mathfrak{q}(\theta)}{\mathbb{E}}\left[\log \mathfrak{q}(\theta)\right]$$

$$\mathfrak{p}(\theta) = \text{Correlated bivariate normal}$$

$$\mathcal{Q} = \{\text{BVN with small, fixed variance}\}$$

What is $\mathfrak{q}^*(\theta) = \underset{\mathfrak{q}\in\mathcal{Q}}{\operatorname{argmin}} \mathrm{KL}\left(\mathfrak{q}(\theta)||\mathfrak{p}(\theta)\right)$?



Sufficently concentrated distributions with constant entropy act like a point mass at the maximum of $\log \mathfrak{p}(\theta)$.

Recall the EM algorithm

$$\text{Observations: } y = (y_1, \ldots, y_N)$$
$$\text{Unknown latent variables: } z = (z_1, \ldots, z_N)$$
$$\text{Unknown global parameter: } \theta \in \mathbb{R}^D. \quad \text{We want: } \hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log p(y|\theta).$$

The EM algorithm alternates between two steps. Starting at an iterate $\hat{\theta}_{(i)}$, repeat until convergence:

The E-step: Compute $Q_{(i)}(\theta) := \underset{p(z|y,\hat{\theta}_{(i)})}{\mathbb{E}} [\log p(y|\theta, z) + \log p(z|\theta)]$

The M-step: Compute the next iterate $\hat{\theta}_{(i+1)} := \underset{\theta}{\operatorname{argmax}}\, Q_{(i)}(\theta)$

The EM algorithm works / is useful when:

- The joint log probability $\log p(y|\theta, z) + \log p(z|\theta)$ is easy to write down
- The posterior $p(z|y, \theta)$ is easy to compute
- The marginalizing integral $p(y|\theta) = \int p(y|\theta, z)p(z|\theta)dz$ is hard

---

Is the EM algorithm VI? **Can you spot the lie?**

The EM algorithm as VI

Let $\mathcal{Q}_z$ denote a family of distributions on z, parameterized by a finite-dimensional parameter η, such that $p(z|\theta, y) \in \mathcal{Q}_z$ for the observed y and all θ.

Exercise: When does $\mathcal{Q}_z$ exist? (Indexed by a finite-dimensional parameter η.)

Let $\mathfrak{q}(z|\hat{\eta}(\theta)) := p(z|\theta, y)$.

In an abuse of notation, write $\eta \in \mathcal{Q}_z$ for $\eta \in \{\eta : \mathfrak{q}(z|\eta) \in \mathcal{Q}_z\}$.

Then:

$$
\begin{aligned}
\log p(y|\theta) &= \log p(y|\theta) + \mathrm{KL}\left(\mathfrak{q}(z|\hat{\eta}(\theta))||p(z|\theta, y)\right) \\
&= \log p(y|\theta) + \operatorname*{argmax}_{\eta \in \mathcal{Q}_z} \left(-\mathrm{KL}\left(\mathfrak{q}(z|\eta)||p(z|\theta, y)\right)\right) && \bigstar \\
&= \operatorname*{argmax}_{\eta \in \mathcal{Q}_z} \left(\log p(y|\theta) - \mathrm{KL}\left(\mathfrak{q}(z|\eta)||p(z|\theta, y)\right)\right) \\
&= \operatorname*{argmax}_{\eta \in \mathcal{Q}_z} \left(\log p(y|\theta) + \mathop{\mathbb{E}}_{\mathfrak{q}(z|\eta)} \left[\log p(z|\theta, y) - \log \mathfrak{q}(z|\eta)\right]\right) \\
&= \operatorname*{argmax}_{\eta \in \mathcal{Q}_z} \left(\mathop{\mathbb{E}}_{\mathfrak{q}(z|\eta)} \left[\log p(y|\theta) \log p(z|\theta, y) - \log \mathfrak{q}(z|\eta)\right]\right) \\
&= \operatorname*{argmax}_{\eta \in \mathcal{Q}_z} \left(\mathop{\mathbb{E}}_{\mathfrak{q}(z|\eta)} \left[\log p(y, z|\theta)\right] + \mathop{\mathbb{E}}_{\mathfrak{q}(z|\eta)} \left[\log \mathfrak{q}(z|\eta)\right]\right) && \bigstar\bigstar
\end{aligned}
$$

The EM algorithm as VI

From the previous slide, the marginal MLE is given by

$$
\begin{aligned}
\hat{\theta} &:= \operatorname*{argmax}_{\theta} \log p(y|\theta) \\
&= \operatorname*{argmax}_{\theta} \operatorname*{argmax}_{\eta \in \mathcal{Q}_z} \left(\log p(y|\theta) - \mathrm{KL}\left(\mathfrak{q}(z|\eta) || p(z|\theta, y) \right) \right) && \bigstar \\
&= \operatorname*{argmax}_{\theta} \operatorname*{argmax}_{\eta \in \mathcal{Q}_z} \left(\underset{\mathfrak{q}(z|\eta)}{\mathbb{E}} \left[\log p(y, z|\theta) \right] + \underset{\mathfrak{q}(z|\eta)}{\mathbb{E}} \left[\log \mathfrak{q}(z|\eta) \right] \right) && \bigstar\bigstar
\end{aligned}
$$

The EM algorithm revisited. Starting at an iterate $\hat{\theta}_{(i)}$:

The E-step:

1. For a fixed $\hat{\theta}_{(i)}$, optimize $\bigstar$ for η. Since only the KL divergence depends on η, the optimum is $\hat{\eta}(\hat{\theta}_{(i)})$, and $\mathfrak{q}(z|\hat{\eta}(\hat{\theta}_{(i)})) = p(z|\hat{\theta}_{(i)}, y)$.
2. Then use $\hat{\eta}(\hat{\theta}_{(i)})$ to compute the expectation in $\bigstar\bigstar$ as a function of θ.

The M-step: Keeping η fixed at $\hat{\eta}(\hat{\theta}_{(i)})$), optimize $\bigstar\bigstar$ as a function of θ to give $\hat{\theta}_{i+1}$. The entropy $\underset{\mathfrak{q}(z|\eta)}{\mathbb{E}} \left[\log \mathfrak{q}(z|\eta) \right]$ does not depend on θ and can be ignored.

$\Rightarrow$ The EM algorithm is coordinate ascent on the objective

$$
f(\theta, \eta) = \log p(y|\theta) - \mathrm{KL}\left(\mathfrak{q}(z|\eta) || p(z|\theta, y) \right).
$$

The EM algorithm as VI

$$\hat{\theta}, \hat{\eta} := \underset{\theta, \eta \in \mathcal{Q}_z}{\operatorname{argmax}} \left(\log p(y|\theta) - \mathrm{KL}\left(\mathfrak{q}(z|\eta) || p(z|\theta, y) \right) \right).$$

The EM algorithm is coordinate ascent on the above objective.

Corrolaries:

- The EM algorithm converges to a local optimum of $\log p(y|\theta)$.
- The EM algorithm is VI, and you don't need to optimize with coordinate ascent.
- If both $p(z|\theta, y)$ and $p(z, y|\theta)$ are easy, then so is $p(y|\theta)$. (This was the lie.)
- If $p(z|\theta, y)$ is intractable, we can now consider different approximating families which may not contain $p(z|\theta, y)$.

Different approximating families: Point masses on z.

Suppose instead of $\mathcal{Q}_z$ we used $\mathcal{Q}_z^{pm}$, a family of constant-entropy near-point mass distributions on z, located at some free parameter η. Then

$$\begin{aligned}
&\underset{\theta,\eta\in\mathcal{Q}_z^{pm}}{\operatorname{argmax}} \left(\log p(y|\theta) - \mathrm{KL}\left(\mathfrak{q}(z|\eta)||p(z|\theta, y)\right)\right) \\
&\quad = \underset{\theta,\eta\in\mathcal{Q}_z^{pm}}{\operatorname{argmax}} \left(\underset{\mathfrak{q}(z|\eta)}{\mathbb{E}} \left[\log p(y, z|\theta)\right] + \underset{\mathfrak{q}(z|\eta)}{\mathbb{E}} \left[\log \mathfrak{q}(z|\eta)\right] \right) \\
&\quad = \underset{\theta,\eta\in\mathcal{Q}_z^{pm}}{\operatorname{argmax}} \left(\underset{\mathfrak{q}(z|\eta)}{\mathbb{E}} \left[\log p(y, z|\theta)\right] \right) = \underset{\theta, z}{\operatorname{argmax}} \log p(y, z|\theta).
\end{aligned}$$

$\Rightarrow$ The Neyman-Scott paradox occurs because point masses are poor approximations for the distribution $p(z|\theta, y)$.

Different approximating families: Point masses on θ.

Let $\mathcal{Q}_\theta^{pm}$ denote a family of constant-entropy near-point mass distributions on θ, located at some free parameter ϑ. Then

$$
\begin{aligned}
&\underset{\theta, \eta \in \mathcal{Q}_z}{\operatorname{argmax}} \left(\log p(y|\theta) - \mathrm{KL}\left(\mathfrak{q}(z|\eta) || p(z|\theta, y) \right) \right) \\
&\quad = \underset{\vartheta \in \mathcal{Q}_\theta^{pm}, \eta \in \mathcal{Q}_z}{\operatorname{argmax}} \left(\underset{\mathfrak{q}(\theta|\vartheta)}{\mathbb{E}} \left[\log p(y|\theta) \right] - \underset{\mathfrak{q}(\theta|\vartheta)}{\mathbb{E}} \left[\log \mathfrak{q}(\theta|\vartheta) \right] - \mathrm{KL}\left(\mathfrak{q}(z|\eta) || p(z|\theta, y) \right) \right) \\
&\quad = \underset{\vartheta \in \mathcal{Q}_\theta^{pm}, \eta \in \mathcal{Q}_z}{\operatorname{argmax}} \mathrm{KL}\left(\mathfrak{q}(\theta|\vartheta) \mathfrak{q}(z|\eta) || \log p(z, \theta|y) \right).
\end{aligned}
$$

$\Rightarrow$ The marginal MLE is a point-mass approximation to the posterior with a uniform prior. It will be a good approximation when $p(\theta|y)$ is approximately a point mass.

Different approximating families.

Suppose we can't compute $p(z|\theta, y)$, and we think that $p(\theta|y)$ may not be well-approximated by a point mass.

Choose some tractable approximating family $\mathfrak{q}(\theta, z|\gamma) \in \mathcal{Q}_{\theta z}$. Then find

$$\hat{\gamma} := \underset{\gamma \in \mathcal{Q}_{\theta z}}{\mathrm{argmin}} \, \mathrm{KL}\left(\mathfrak{q}(\theta, z|\gamma) || p(\theta, z|y)\right).$$

Now we're doing "Variational Bayes".

Some common approximating families:

- Factorizing families, e.g. $\mathfrak{q}(\theta, z|\gamma) = \mathfrak{q}(\theta|\gamma)\mathfrak{q}(z|\gamma)$. These families model the posteriors as independent.
  - For historical reasons, the factorizing assumption is known as a **mean-field approximation**.
- Factorizing families + an exponential family assumption.
- Normal approximations (possibly after an invertible unconstraining transformation): $\mathfrak{q}(\theta, z|\gamma) = \mathcal{N}(\theta, z|\gamma)$.
- Independent normal approximations. This is used by a lot of "black-box VI" methods [Ranganath et al., 2014, Kucukelbir et al., 2017].

Conclusions

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Eric Maris, and Marcel AJ van Gerven. Wasserstein variational inference. *arXiv preprint arXiv:1805.11284*, 2018.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22 (1):79–86, 1951.

Yingzhen Li and Richard E Turner. Variational inference with rényi divergence. *stat*, 1050:6, 2016.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.